Data and Statistical Analyses: Tips and Tricks

Daniel Normolle, PhD
University of Pittsburgh Department of Biostatistics
UPMC Hillman Cancer Center

February 21, 2019

Society for Immunotherapy of Cancer

I Hate p-Values!

Society for Immunotherapy of Cancer

I Hate p-Values

Karl Pearson

Society for Immunotherapy of Cancer

I Hate p-Values

Karl Pearson

George Udny Yule

Society for Immunotherapy of Cancer

I Hate p-Values

Karl Pearson

George Udny Yule

William Gossett (Student)

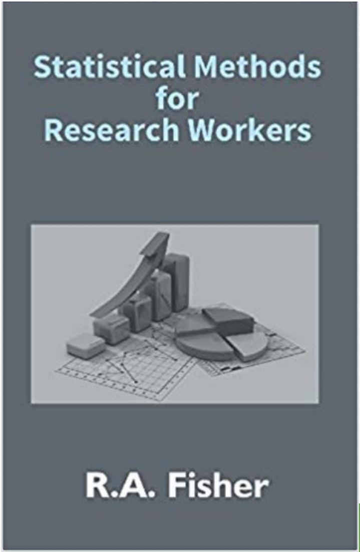
Society for Immunotherapy of Cancer

I Hate p-Values

Ronald Aylmer Fisher

Society for Immunotherapy of Cancer

What's the Story with $p < 0.05$?



Society for Immunotherapy of Cancer

What's the Story with $p < 0.05$?

F - Distribution ($\alpha = 0.01$ in the Right Tail)

Denominator Degrees of Freedom df_2 \ df_1 (Numerator Degrees of Freedom)	1	2	3	4	5	6	7	8	9
1	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158
6	13.745	10.925	9.7795	9.1483	8.7459	8.4661	8.2600	8.1017	7.9761
7	12.246	9.5466	8.4513	7.8466	7.4604	7.1914	6.9928	6.8400	6.7188
8	11.259	8.6491	7.5910	7.0061	6.6318	6.3707	6.1776	6.0289	5.9106
9	10.561	8.0215	6.9919	6.4221	6.0569	5.8018	5.6129	5.4671	5.3511
10	10.044	7.5594	6.5523	5.9943	5.6363	5.3858	5.2001	5.0567	4.9424
11	9.6460	7.2057	6.2167	5.6683	5.3160	5.0692	4.8861	4.7445	4.6315
12	9.3302	6.9266	5.9525	5.4120	5.0643	4.8206	4.6395	4.4994	4.3875
13	9.0738	6.7010	5.7394	5.2053	4.8616	4.6204	4.4410	4.3021	4.1911
14	8.8616	6.5149	5.5639	5.0354	4.6950	4.4558	4.2779	4.1399	4.0297
15	8.6831	6.3589	5.4170	4.8932	4.5556	4.3183	4.1415	4.0045	3.8948
16	8.5310	6.2262	5.2922	4.7726	4.4374	4.2016	4.0259	3.8896	3.7804
17	8.3997	6.1121	5.1850	4.6690	4.3359	4.1015	3.9267	3.7910	3.6822
18	8.2854	6.0129	5.0919	4.5790	4.2479	4.0146	3.8406	3.7054	3.5971
19	8.1849	5.9259	5.0103	4.5003	4.1708	3.9386	3.7653	3.6305	3.5225
20	8.0960	5.8489	4.9382	4.4307	4.1027	3.8714	3.6987	3.5644	3.4567
21	8.0166	5.7804	4.8740	4.3688	4.0421	3.8117	3.6396	3.5056	3.3981
22	7.9454	5.7190	4.8166	4.3134	3.9880	3.7583	3.5867	3.4530	3.3458
23	7.8811	5.6637	4.7649	4.2636	3.9392	3.7102	3.5390	3.4057	3.2986
24	7.8229	5.6136	4.7181	4.2184	3.8951	3.6667	3.4959	3.3629	3.2560
25	7.7698	5.5680	4.6755	4.1774	3.8550	3.6272	3.4568	3.3239	3.2172
26	7.7213	5.5263	4.6366	4.1400	3.8183	3.5911	3.4210	3.2884	3.1818
27	7.6767	5.4881	4.6009	4.1056	3.7848	3.5580	3.3882	3.2558	3.1494
28	7.6356	5.4529	4.5681	4.0740	3.7539	3.5276	3.3581	3.2259	3.1195
29	7.5977	5.4204	4.5378	4.0449	3.7254	3.4995	3.3303	3.1982	3.0920
30	7.5625	5.3903	4.5097	4.0179	3.6990	3.4735	3.3045	3.1726	3.0665
40	7.3141	5.1785	4.3126	3.8283	3.5138	3.2910	3.1238	2.9930	2.8876
60	7.0771	4.9774	4.1259	3.6490	3.3389	3.1187	2.9530	2.8233	2.7185
120	6.8509	4.7865	3.9491	3.4795	3.1735	2.9559	2.7918	2.6629	2.5586
∞	6.6349	4.6052	3.7816	3.3192	3.0173	2.8020	2.6393	2.5113	2.4073

Society for Immunotherapy of Cancer

Multiple Comparions

- If you report many p-values, you increase your *family-wise* error rate
- How many tests did you run, and how many did you report?
- Adjust. . .
  - Bonferroni
  - Tukey
  - Stepdown
- These will reduce your FWER, but impact your power. You could not adjust. . .
  - Say how many tests you ran
  - Don't adjust the tests
  - Let the reader decide

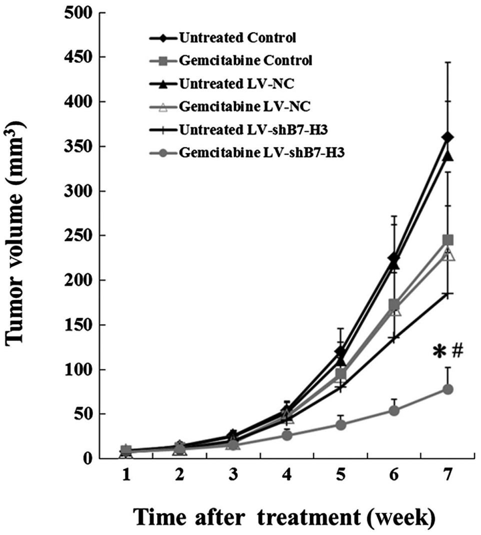
Society for Immunotherapy of Cancer

Repeated Measures Are Your Friends

Society for Immunotherapy of Cancer

What Is The Point Here?

- How many animals are in each curve?
- How many animals are being lost to ethical sacrifice?
- Are we just interested in the comparison at one time point?
- Why are the confidence intervals getting bigger?



Society for Immunotherapy of Cancer

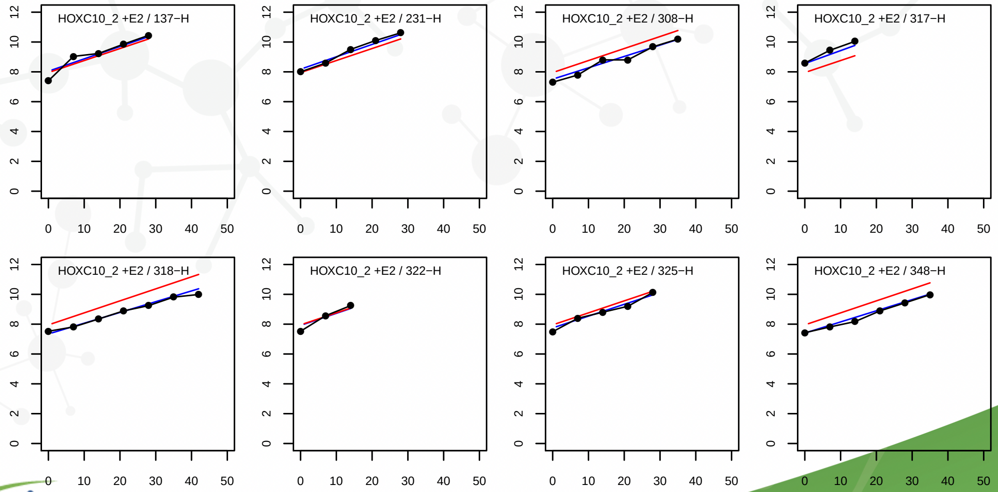
A Better Idea

- Linearize exponential growth with the log transform
- Fit a line to each animal
- The slope of the line is the tumor growth rate
- Compare growth rates between treatment groups

- No bias induced by ethical sacrifice
- Increase statistical power
- Use *all* the data!

Society for Immunotherapy of Cancer

Fit The Animals, Average the Fits



Society for Immunotherapy of Cancer

Repeated Measures in Xenograft Experiments

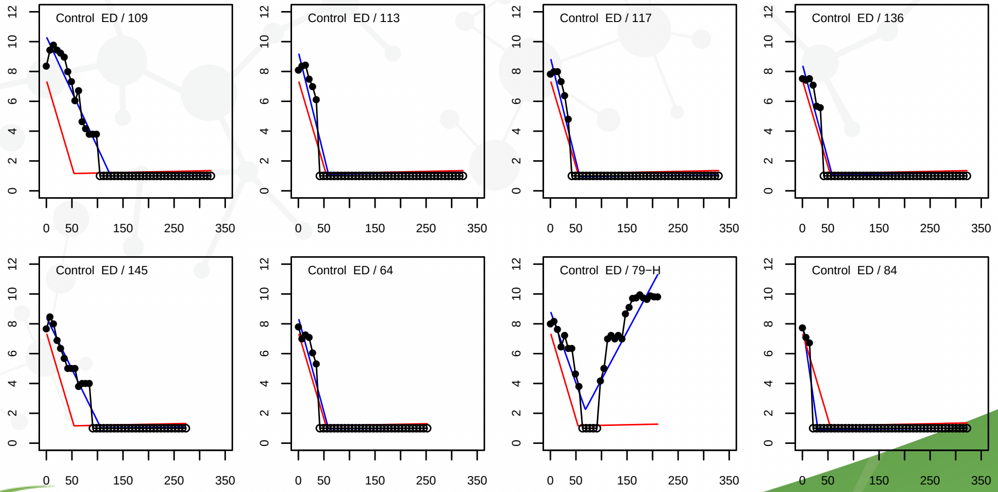
- Random slopes and intercepts (mixed effects) model does the analysis above in one model

$$\log_2(V_{ij}(t)) = \underbrace{\alpha_i + \beta_i t}_{\text{Pop.}} + \underbrace{a_{ij} + b_{ij} t}_{\text{Animal}} + \underbrace{e_{ijt}}_{\text{Noise}}$$

- Compare βs using Wald-type tests
- Software
  - SAS Proc Mixed
  - R function lmer() in package lmerTest
  - Python function mixedlm() in package statsmodels

Society for Immunotherapy of Cancer

"Broken Stick" Model

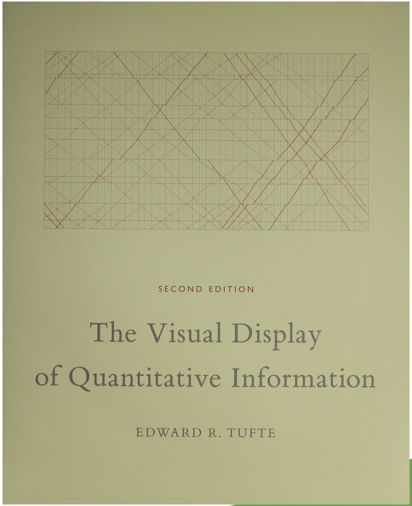


I Agree With Leslie

Society for Immunotherapy of Cancer

A Good Book

Not to be confused with *The* Good Book



Society for Immunotherapy of Cancer

What Is All This?

Society for Immunotherapy of Cancer

What Is *Data Science*?

- Term was coined by Bell Labs Statistician Bill Cleveland in 1991
- He used it to compare (favorably) a proposed course of training to mine:
  - Explicit training in computer programming
  - Training in data wrangling
  - Emphasis on collaborative research
- Some data science programs will have more emphasis on advanced applied computing (e.g., how to use a Hadoop Cluster) and machine learning techniques

Society for Immunotherapy of Cancer

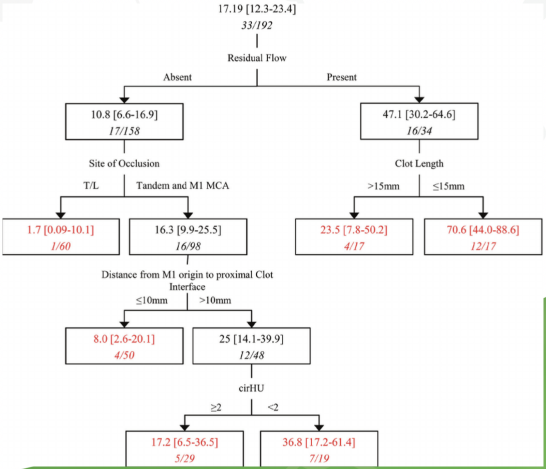
What Is *Machine Learning*?

- Machine learning emphasizes Predictive Modeling, not inference (no [redacted] p-values!)
- It includes many traditional statistical methods: regression, logistic regression, principal components, regularized regression
- It also includes "black box" predictive modeling methods that many statisticians know about, too: random forests, neural nets, ensemble modeling, support vector machines
- These methods can offer excellent predictive performance in cases of significant nonlinearity or very many predictors

sitc
Society for Immunotherapy of Cancer

What Is *Random Forests*?

- It's souped up Recursive Partitioning ⇒
- Regression or Classification
- Recursive Partitioning has a reputation for over-fitting and sensitivity to small permutations in the data
- So, Leo Breiman fixed it



sitc
Society for Immunotherapy of Cancer

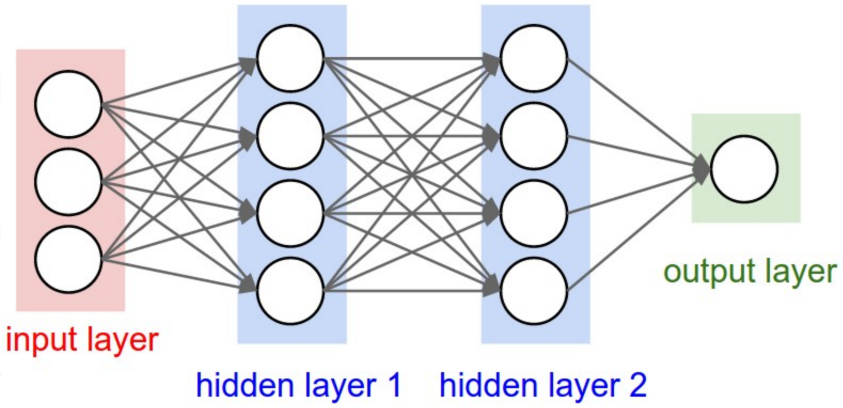
What Is *Random Forests*?

- Do the following many (>tens of thousands of) times:
  - Take a bootstrap sample (with replacement) of the inputs and outputs in the training set
  - Take a random sample of the inputs ($\sqrt{p}$)
  - Fit the bootstrap sample inputs to the outputs using the random sample of the inputs
- Average the predictions of all those trees
- It works better. You can demonstrate it.

Society for Immunotherapy of Cancer

What Is A *Neural Net*?

- It's a bunch of logistic regressions that feed into each other to ultimately produce a prediction



What Is A *Neural Net*?

- Estimation of the models' parameters are done by *backpropagation*
- The estimation can be time consuming, but the resulting predictions can be made really fast
- There are implementations in R, Python, SAS and Matlab (and elsewhere, I'm sure)

sitc
Society for Immunotherapy of Cancer

What Is *Deep Learning*?

- A neural net with more layers

Society for Immunotherapy of Cancer

What Is *Deep Learning*?

- A neural net with more layers
- It's getting deep in here. . .

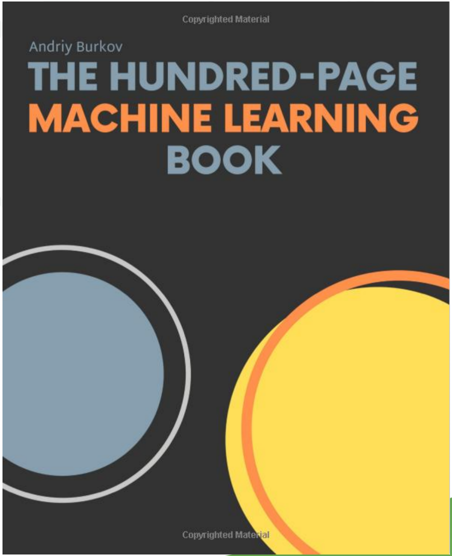
Society for Immunotherapy of Cancer

What Is An *Ensemble Model*?

- Run all the methods above to build predictive models
- Average them

Society for Immunotherapy of Cancer

Another Good Book



Society for Immunotherapy of Cancer

What Is *Artifical Intelligence*?

Society for Immunotherapy of Cancer

What Is *Artifical Intelligence*?

Who the [redacted] knows?

Society for Immunotherapy of Cancer

Optimal Dichotimization

Society for Immunotherapy of Cancer

There Is An Optimal Cutpoint. Just Sayin'

Expected Cost of a Decision

- Assign a unit to Population 1 or 2 based on a continuous random variable x
- The cost of correct classification is 0
- $C_1 > 0$ is the cost of assigning a member of Population 1 to Population 2
- $C_2 > 0$ is the cost of assigning a member of Population 2 to Population 1
- It is also possible to generalize to the case where there is a non-zero cost of a correct decision
- π is the prevalence of population 1 in the mixture
- Assume WLOG that the rule is: Assign member to Population 1 if $x \leqslant c$

sitc
Society for Immunotherapy of Cancer

Optimal Cutpoint

Minimize the Expected Cost

- The expected cost of this rule is:

$$
\begin{aligned}
E(c) &= \pi C_1 \int_{-\infty}^{c} f_1(x)dx + (1-\pi)C_2 \int_{c}^{\infty} f_2(x)dx \\
&= \pi C_1 F_1(c) + (1-\pi)C_2(1 - F_2(c))
\end{aligned}
$$

- To find the optimal value of c, set

$$
\frac{d}{dc}E(c) = 0
$$

Society for Immunotherapy of Cancer

Optimal Cutpoint

Minimize the Expected Cost

- Then,

$$\begin{aligned}
\frac{d}{dc}E(c) &= \pi C_1 f_1(c) - (1-\pi)C_2 f_2(c) = 0 \Rightarrow \\
\pi C_1 f_1(c) &= (1-\pi)C_2 f_2(c) \Rightarrow \\
\frac{\pi C_1}{(1-\pi)C_2} &= \frac{f_2(c)}{f_1(c)}
\end{aligned}$$

- This is the *Bayes minimum risk decision rule*

Society for Immunotherapy of Cancer

Optimal Cutpoint

- Need π, C_1, C_2, f_1 and f_2
- f_1 and f_2 can be estimated from a sample
- π is the prevalence in the population, not in the sample
- It's hard to come to a consensus on C_1 and C_2,
- Especially when one of the costs involves death

Society for Immunotherapy of Cancer

What is Bayesian Statistics, Anyway?

Society for Immunotherapy of Cancer

Why Consider Bayesian Methods?

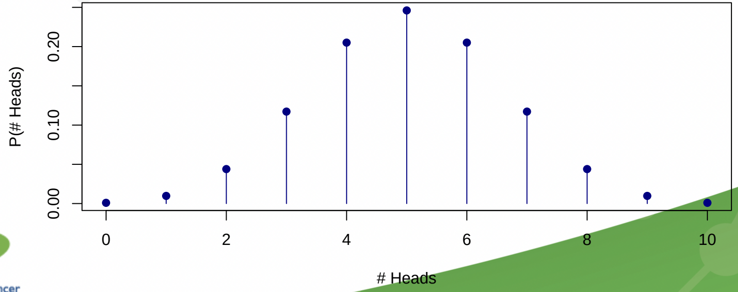
- Explicitly incorporate information outside the experiment into the data analysis
- Great flexibility in probability models
- Allow multiple looks at the data without penalty for multiple tests
- Monitoring clinical trials
- Model-based Phase I designs
- Response- and biomarker-adaptive clinical trials

Society for Immunotherapy of Cancer

The Frequentist Model of the Universe

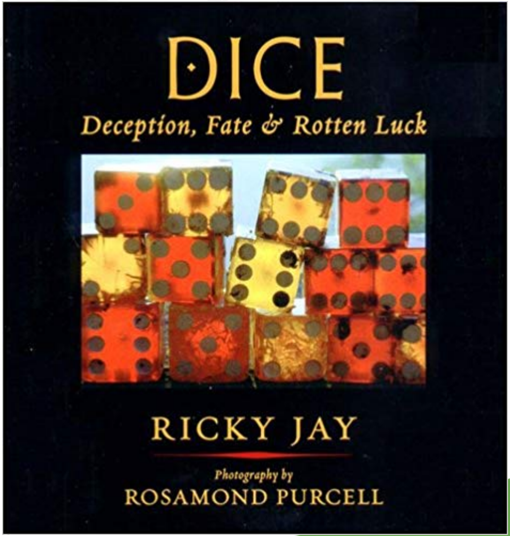
- Repeated sampling of a random variable, $\{X_1, X_2, X_3 \ldots\}$ is described by a *probability distribution function*, indexed by *parameters*, e.g., $\mu, \sigma, \pi \ldots$
- These parameters are *unobservable* and *fixed*
- Example: number of heads in $n = 10$ coin tosses, where $\pi = 1/2$:

$$P(x \text{ heads in } n \text{ tosses} \mid \pi) = \binom{x}{n} \pi^x (1 - \pi)^{n-x}$$



Society for Immunotherapy of Cancer

Yet Another



sitc
Society for Immunotherapy of Cancer

The Bayesian Model of the Universe

- Repeated sampling of a random variable, $\{X_1, X_2, X_3 \ldots\}$ is described by a *probability distribution function*, indexed by *parameters*, e.g., $\mu, \sigma, \pi \ldots$
- These parameters are *unobservable* and *random*
- The parameters, being random variables themselves, have a probability distribution function themselves, which is called a *prior distribution*
- Example: number of heads in ten coin tosses:

$$
\begin{aligned}
\mathrm{P}(x \text{ heads in } n \text{ tosses} | \pi) &= \binom{x}{n} \pi^x (1-\pi)^{n-x} \\
\pi &\sim \mathrm{Beta}(\alpha, \beta)
\end{aligned}
$$

- α and β are called the *prior parameters* or *metaparameters*. They are set by the analyst.

Society for Immunotherapy of Cancer

The Bayesian Model of the Universe

- Posterior $\leftarrow$ Prior + Data
- The endpoint of a Bayesian analysis is a statement about the posterior distribution function (i.e., *after* the data have been observed) of the parameters of the probability function of the data
  - Frequentist: "A 95% confidence interval for π is $(0.25, 0.45)$"
  - Bayesian: "There is a 95% probability that π is between 0.25 and 0.45"
- Note that these are *not* the same
- The Bayesian model is convenient in continuous (or frequent) monitoring of clinical trials because hypothesis testing is not involved, so there is no inflation of the Type I error from multiple testing

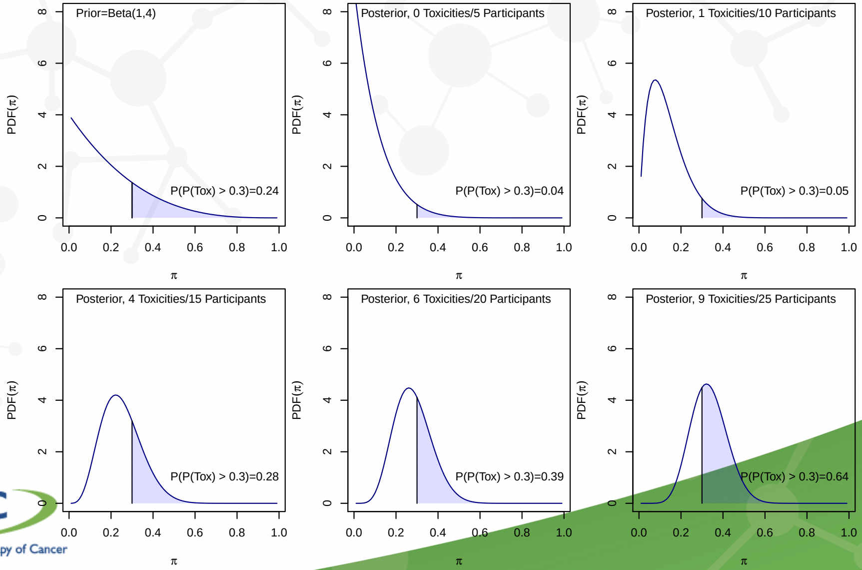
sitc
Society for Immunotherapy of Cancer

Example: Toxicity Monitoring

- "Accrual will be halted and the trial will be reëvaluated if $P(P(\text{Toxicity}) > 0.3) > 0.6$"
- We have a belief, not very strong, that the probability of toxicity, π, is around 0.2, so we choose a Beta(1,4) prior, which has the weight of $1 + 4 = 5$ observations.

Society for Immunotherapy of Cancer

Example: Toxicity Monitoring



sitc
Society for Immunotherapy of Cancer

Bayesian Statistics in Medical Device Trials

- FDA defines a medical *device* as as any product that does not achieve its purposes by chemical action or metabolization
- Many types of devices:

Contact lenses	Breast implants	MRI machines
Surgical instruments	Hip replacements	Thermometers
Artificial hearts	Hearing aids	Latex gloves
Surgical stents	Diagnostic tests	Defibrillators

- Many (1,000s) small manufacturers
- Average life length of a device is two years
- Registration system for medical devices differs from that of drugs

Society for Immunotherapy of Cancer

Bayesian Statistics in Medical Device Trials

- There are three classes of medical devices:
  - Class 1: Low risk, requiring only general controls (examples: adhesive bandage, sunglasses)
  - Class 2: Moderate risk, requiring general controls and special controls (examples: syringe, surgical mask, powered wheelchair)
  - Class 3: High risk, requiring general controls and pre-marketing approval (examples: heart valves, implantable neuromuscular stimulator)

Bayesian Statistics in Medical Device Trials

- General controls:
  - Adulteration and misbranding
  - Quality systems
  - Labeling
  - Medical device reporting
  - Electronic Establishment Registration
  - Electronic Device Listing
  - Premarket Notification [510(k)]

Society for Immunotherapy of Cancer

Bayesian Statistics in Medical Device Trials

- Special controls:
  - Guidelines (e.g., Glove Manual)
  - Mandatory Performance Standard
  - Recommendations or Other Actions
  - Special Labeling, specified in detail in 21 CFR 882

Bayesian Statistics in Medical Device Trials

- Pre-marketing approval:
  - Requires negotiation with FDA
  - Standards differ depending on the similarity of the new devices to existing devices already marketed
  - May or may not require a clinical trial

Society for Immunotherapy of Cancer

Bayesian Statistics in Medical Device Trials

- Much prior information on similar devices
- For registration trials, FDA requires data-derived priors
- FDA negotiates with sponsor before registration trial on what constitutes valid prior data
- Prior data can be proprietary or publicly available
- Simulations are used to demonstrate operating characteristics
- Required sample sizes can be significantly decreased
- Sometimes use formal economic risk/benefit analysis

Society for Immunotherapy of Cancer

Bayesian Medical Device Trial Example

- TherOx Downstream Aqueous Oxygen System
- Device used to deliver superoxygenated blood to patient's heart after MI
- FDA agreed to single randomized, controlled pivotal trial, AMIHOT I [1]

Society for Immunotherapy of Cancer

[1] No, I did not make this up.

AMIHOT I

- 289 patients from 23 centers, randomized 1:1
- Trial designed for noninferiority of safety endpoint (death, stroke, etc.)
- Trial designed for superiority of efficacy endpoint (infarct size at 14 days, WMSI at three months)
- Safety endpoint succeeded: 9/134 treated versus 7/135 control SAEs, $p < 0.022$
- All three efficacy endpoints failed, all $p > 0.24$
- Post-hoc analysis showed efficacy in anterior MI subset within 6 hours of MI only, all $p < 0.04$, based on 100 patients

Society for Immunotherapy of Cancer

AMIHOT II

- Since the results of AMIHOT I weren't very hot, the sponsors proposed AMIHOT II
- 317 patients from 22 centers, 2.8:1
- Same endpoints as AMIHOT I, anterior MI patients < 6 hours out only
- Safety goal: show $P(\pi_T < \pi_C + 0.06 \mid \text{data and prior}) > 0.95$
- Hierarchical Bayesian model using all AMIHOT I data

sitc
Society for Immunotherapy of Cancer

AMIHOT II

- Hierarchical Bayesian models for safety and efficacy using all AMIHOT I data used four subgroups:
  - Non-anterior MI, >6 Hours (AMIHOT I only)
  - Non-anterior MI, $\leqslant 6$ Hours (AMIHOT I only)
  - Anterior MI, >6 Hours (AMIHOT I only)
  - Anterior MI, $\leqslant 6$ Hours (AMIHOT I & AMIHOT II)
- Safety model is specified on next slide as an example:

sitc
Society for Immunotherapy of Cancer

AMIHOT II Hierarchical Bayesian Model for Safety

- Let $i = 1, 2$ index study (AIHOT I or II)
- Let $j = 1, 2, 3, 4$ index subgroup (as above)
- Let r be a safety-related event, C be control and T be treatment
- Population model:

$$\begin{aligned} r_{ij}^C &\sim \text{Binomial}(n_{ij}^C, \pi_{ij}^C) \\ r_{ij}^T &\sim \text{Binomial}(n_{ij}^T, \pi_{ij}^T) \end{aligned}$$

- Parameter model:

$$\begin{aligned} \lambda_{ij}^C &= \text{logit}(\pi_{ij}^C) \\ \lambda_{ij}^C &= \mu_0 + \omega_j^C + \gamma_i^C \\ \pi_{ij}^T &= \pi_{ij}^C + \delta_0 + \omega_j^T + \gamma_i^T I(0, 1) \end{aligned}$$

$$\omega_j^C \sim \text{Normal}(0, \phi_\omega^2),\ \gamma_j^C \sim \text{Normal}(0, \phi_\gamma^2),$$
$$\omega_j^T \sim \text{Normal}(0, \tau_\omega^2),\ \gamma_j^T \sim \text{Normal}(0, \tau_\gamma^2)$$

Society for Immunotherapy of Cancer

Three Rules From Wise Men

Society for Immunotherapy of Cancer

Nelson Algren (1909-81)

- Writer
- *The Man With The Golden Arm*
- *A Walk On The Wild Side*
- Simone de Beauvoir's lover!
- 500 page FBI dossier

Nelson Algren's Three Rules

- Never eat at a place named "Mom's"

Society for Immunotherapy of Cancer

Nelson Algren's Three Rules

- Never eat at a place named "Mom's"
- Never play cards with a man named "Doc"

Society for Immunotherapy of Cancer

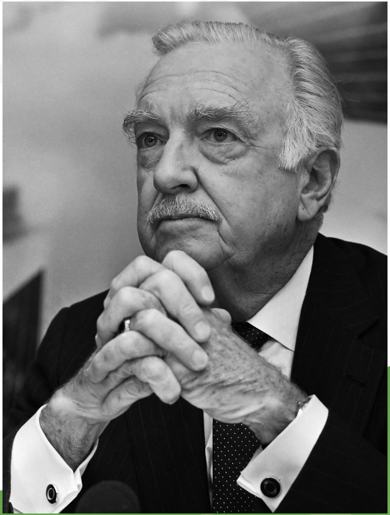
Nelson Algren’s Three Rules

- Never eat at a place named “Mom’s”
- Never play cards with a man named “Doc”
- Never sleep with somebody who’s got more troubles than you have

Society for Immunotherapy of Cancer

Walter Cronkite (1916-2009)

- CBS Evening News anchor for 19 years
- "The Most Trusted Man In America"
- Anchored all the moon landings
- If I've lost Cronkite, I've lost Middle America–Lyndon Johnson



Society for Immunotherapy of Cancer

Walter Cronkite's Three Rules for Old Men

- Never pass up a free drink

Society for Immunotherapy of Cancer

Walter Cronkite's Three Rules for Old Men

- Never pass up a free drink
- Never [redacted]

Society for Immunotherapy of Cancer

Walter Cronkite's Three Rules for Old Men

- Never pass up a free drink
- Never [redacted]
- Never [redacted]

Society for Immunotherapy of Cancer

The End!

Society for Immunotherapy of Cancer